On the Semantics of Noun Compounds

Roxana Girju *

Computer Science Department, Baylor University, Waco, Texas 76798, USA

Dan Moldovan, Marta Tatu, Daniel Antohe

Human Language Technology Research Institute, University of Texas at Dallas, Richardson, TX, USA.

Abstract

This paper provides new insights on the semantic characteristics of two and three noun compounds. An analysis is performed using two sets of semantic classification categories: a list of 8 prepositional paraphrases previously proposed by Lauer (Lauer 1995) and a new set of 35 semantic relations introduced by us. We show the distribution of these semantic categories on a corpus of noun compounds and present several models for the bracketing and the semantic classification of noun compounds. The results are compared against state-of-the-art models reported in the literature. We also demonstrate the applicability of noun compounds semantics to the question answering problem.

Key words: noun compounds, computational lexical semantics

1 Introduction

The semantic interpretation of noun compounds (NCs) deals with the detection and semantic classification of the relations between noun constituents. The problem is complex and has been studied intensively in linguistics, psycholinguistics, philosophy, and computational linguistics for a long time. There are several reasons that make this task difficult. (a) NCs have implicit semantic relations; for example, "*spoon handle*" encodes a PART-WHOLE relation. (b) NCs'

* Corresponding author.
Email addresses: `girju@cs.baylor.edu` (Roxana Girju), `moldovan@utdallas.edu`, {`marta`, `dantohe`}`@hlt.utdallas.edu` (Dan Moldovan, Marta Tatu, Daniel Antohe).

Preprint submitted to Elsevier Science *5 June 2004*

interpretation is knowledge intensive and can be idiosyncratic. For example, to correctly interpret "*GM car*" one has to know that GM is a car-producing company. (c) There can be more than one semantic relation encapsulated in a pair of nouns. For example, "*Texas city*" can be tagged as a part-whole relation as well as a location relation. (e) The interpretation of NCs can be highly context-dependent. For example, "apple juice seat" can be defined as "seat with apple juice on the table in front of it" (cf. (Downing 1977)).

Although researchers ((Jespersen 1954), (Downing 1977)) argued that noun compounds encode an infinite set of semantic relations, many agree ((Levi 1978), (Finin 1980)) there is a limited number of relations that occur with high frequency in noun compounds. However, the number and the level of abstraction of these frequently used semantic categories are not agreed upon. They can vary from a few prepositional paraphrases (Lauer 1995) to hundreds and even thousands more specific semantic relations (Finin 1980). The more abstract the categories, the more noun compounds are covered, but also the more room for variation as to which category a compound should be assigned to. Lauer (Lauer 1995), for example, considers eight prepositional paraphrases as semantic classification categories: *of, for, with, in, on, at, about*, and *from*. According to this classification, the noun compound "*birds sanctuary*", for instance, can be classified both as "*sanctuary* **of** *birds*" and "*sanctuary* **for** *birds*". The main problem with these abstract categories is that much of the meaning of individual compounds is lost, and sometimes there is no way to decide whether a form is derived from one category or another.

On the other hand, lists of very specific semantic relations are difficult to build as they usually contain a very large number of predicates, such as the list of all possible verbs that can link the noun constituents. Finin (Finin 1980), for example, uses semantic categories such as "**dissolved in**" to build interpretations of compounds like "*salt water*" and "*sugar water*". Although, there were several proposals of possible large sets of semantic relations, there has been no attempt to map one set to another, and more importantly, to define the most appropriate level of abstraction for the interpretation of compounds in general, or for a specific application in particular.

Due to the recursiveness of compounding (Selkirk 1982), much of the semantics of two-word noun compounds applies to multi-word compounds. However, the interpretation problem becomes significantly more complicated for larger noun sequences, such as three noun compounds since both the modifier and the head nouns can form noun compounds generating *structural ambiguities*. This task is called *bracketing* or *attachment* and is the first step in interpreting multi-word noun compounds. Choosing the most probable binary bracketing for a given noun sequence represents a difficult task as attachments are not syntactically, but semantically governed. Consider, for example, the following noun compounds: (1) *((consumer confidence) survey)*, (2) *(state (gasoline tax))*, and

(3) *(car (radio equipment)) or ((car radio) equipment).* The noun compound (1) is left-bracketed, while the noun compound (2) is right-bracketed. There are also situations such as (3) in which both left and right-branching solutions are possible. Sometimes, the disambiguation is provided only by the context.

The automatic interpretation of noun compounds is a difficult task for both unsupervised and supervised approaches. Currently, the best-performing NC interpretation methods in computational linguistics focus only on two-word noun compounds and rely either on rather ad-hoc, domain-specific, hand-coded semantic taxonomies, or on statistical models on large collections of unlabeled data. Recent results have shown that symbolic NC interpretation systems using Machine Learning techniques coupled with a large lexical hierarchy perform with very good accuracy, but they are most of the time tailored to a specific domain (Rosario and Hearst 2001), (Rosario et al 2002).

The majority of the corpus statistics approaches to the noun compound interpretation collect statistics on the occurrence frequency of the noun constituents and use them in a probabilistic model ((Resnik 1993), (Lauer 1995), (Pustejovsky et al. 1993), (Lapata 2000), (Lapata and Keller 2004)). The problem is that most noun compounds are rare and thus, statistics on such infrequent instances lead in general to unreliable estimates of probabilities. More recently, Lapata and Keller (Lapata and Keller 2004) showed that simple unsupervised models applied to the noun compound interpretation task perform significantly better when the n-gram frequencies are obtained from the web (accuracy of 55.71% on Altavista), rather than from a large standard corpus. However, although the web-based solution might overcome the data sparseness problem, the probabilistic models are limited by the lack of linguistic information. Most of the time the probabilities are computed on lexical items with or without inflected forms. This simplistic approach introduces a number of ambiguities ranging from syntactic and structural, to semantic.

In this paper we describe various domain independent models that use supervised machine learning techniques and a set of linguistic features. The main feature of these models is the use of the word sense disambiguation information of the noun constituents extracted based on their surrounding context. We focus only on two and three-noun compositional compounds, ie. those whose meaning can be derived from the meaning of the constituent nouns (eg, "*door knob*"), and tackle both the bracketing and the interpretation tasks. However, we check if the constructions are lexicalized (non-compositional), ie the meaning is a matter of convention (eg, "*soap opera*"), but only for statistical purposes. We present empirical observations on the distribution of a core set of semantic relations in noun compounds and provide a mapping between two sets of semantic classification categories. The noun compound interpretation system has been tested on a list of 8 general prepositional paraphrases (Lauer 1995) and a list of 35 semantic relations (Moldovan and Girju 2003).

We also compare our results for bracketing and interpretation tasks against two baselines and against two state-of-the-art interpretation systems.

The paper is organized as follows. Section 2 presents the general approach for the interpretation of noun compounds and lists the semantic categories used along with observations regarding the distribution of these semantic categories in the corpus. Sections 3 and 4 present models and results for the interpretation of two, respectively three noun compounds. Section 5 illustrates the applicability of noun compound interpretation to question answering. Finally, some conclusions are offered in Section 6.

2 Approach

We approach the problem top-down, namely identify first the characteristics or feature vectors of noun compounds, then develop models for their semantic classification. This is in contrast to our prior approach (Girju et al 2003a) where we studied one semantic relation at a time, and learned constraints to identify only that relation. The distribution of noun compound semantic relations in a corpus is analyzed shedding some light on the resulting *semantic spaces*. We define a semantic space as the set of relations encoded by noun compounds. We aim at uncovering the general aspects that govern the semantics of noun compounds, and thus delineate the semantic space within different sets of semantic classification categories. These feature vectors are then employed in various leaning models.

2.1 Lists of Semantic Classification Relations

In this paper we consider two sets of semantic classification categories for the noun compounds interpretation. The first is the Lauer's list of 8 prepositional paraphrases presented in Section 1 and the second is a list of 35 semantic relations identified by us after many iterations over a period of time (Moldovan and Girju 2003). This list, presented in Table 3 along with examples, is general enough to cover a large majority of text semantics while keeping the number of the semantic relations to a manageable number.

We selected these sets as they are of different size and contain semantic classification categories at different levels of abstraction. Lauer's list is more abstract and, thus capable of encoding a large number of noun compound instances found in a corpus, while our list contains finer grain semantic categories. We show below the coverage of these semantic lists on a fairly large corpus, how well they solve the interpretation problem, and the mapping from one list to another.

In order to devise an automatic method for the detection of semantic relations in noun compounds, we analyzed the semantic behavior of these constructions on a large, domain independent corpora of examples. Our intention is to answer questions like: (1) *Given a set of semantic classification relations and a corpus of examples, what is the core subset frequently encoded by noun compounds?* otherwise said, *Is there a subset of preferred meanings?*, (2) *What is their distribution on a large corpus?*, (3) *Are there semantic relations that are not allowed in noun compounds?*, (4) *How well can noun compounds be paraphrased with prepositional paraphrases, and respectively with more specific semantic relations?*, and (5) *How many NCs are lexicalized?*

The data

For each type of noun compounds considered, a training corpus was assembled from two sources: Wall Street Journal articles from TREC-9, and eXtended WordNet glosses (XWN 2.0) (http://xwn.hlt.utdallas.edu). We used XWN since all its glosses are syntactically parsed and words are semantically disambiguated which saved us a considerable amount of time. Table 1 shows the number of randomly selected sentences from each text collection and the corresponding number of instances of annotated pairs after the inter-annotator agreement. The annotation of each example consists of specifying its feature vector and the most appropriate interpretation based on: (1) the list of 35 semantic relations (35 SRs) (Table 3) and (2) the Lauer's list of 8 prepositional paraphrases (8 PPs) (cf. (Lauer 1995)).

Since we wanted to compare our approach with two state-of-the-art unsupervised probabilistic approaches, we selected as test sets those randomly selected by Lauer from Grolier Encyclopedia for each type of noun compounds: 282 noun-noun pairs for 2-noun compounds, and 244 three noun instances for the 3-noun compounds. However, the training and test sets for each type of noun compound have different distributions. Thus, we further shuffled the training and test corpora and randomly split them again maintaining the same ratio. We call these training and test sets *unshuffled*, and respectively *shuffled*.

Collection	**2-noun compounds**		**3-noun compounds**
	WSJ	**XWN**	**WSJ**
Number of sentences	3,217	5,672	49,208
Number of annotated training instances after agreement	2,606	1,379	362

Table 1

Number of sentences and training noun compound instances after agreement selected from each text collection considered.

Corpus Annotation and Inter-annotator Agreement

Two PhD students in Computational Semantics have annotated separately all the noun compounds in the training corpora. Sequences of two and three nouns were extracted from syntactically parsed sentences (Charniak's parser - (Charniak 2000)) using the Lauer's heuristic (Lauer 1995) (for XWN we used the gold parse trees). The heuristic looks for consecutive nouns of size two, and respectively three, that are neither preceded nor succeeded by a noun. In order to eliminate the wrong instances selected by the heuristic, the annotators were provided with the sentence in which the nouns occurred and were asked to manually check the noun compounds. The remaining NCs were tagged with their corresponding WordNet senses if found in WordNet (360 instances were not found in WordNet or the correct sense was missing) and semantic classification relations. Whenever the annotators found an example encoding a semantic relation, and respectively a prepositional paraphrase, other than those provided or they didn't know what interpretation to give, they had to tag it as "OTHERS-SR", and respectively "OTHERS-PP". Besides the type of relation, the annotators were asked to provide information about the order of the modifier and the head nouns in a noun - noun pair if applicable. For instance, in "*honey bee*"-MAKE/PRODUCE the product *honey* is followed by the producer *bee*, while in "*GM car*"-MAKE/PRODUCE/r the order is reversed (r means reversed). On average, 34% of the noun - noun training examples, and respectively 24% of the three noun compound instances had at least a noun - noun pair in reverse order.

Most of the time, one instance was tagged with one semantic relation, and respectively prepositional paraphrase, but there were also situations in which an example could belong to more than one relation in the same context. For example, "*Texas city*" was tagged as a PART-WHOLE/PLACE-AREA relation and as a LOCATION relation, and respectively as "*city of Texas*", "*city from Texas*", and "*city in Texas*". Overall, for noun - noun compounds there were 608 instances tagged with more than one semantic relation, and respectively almost all paraphrasable instances were tagged with more than one prepositional paraphrase. Moreover, the annotators were asked to indicate if the instance was lexicalized or not. They found that 30% of the 2-noun compounds were lexicalized, from which 18% were proper names.

For three noun compounds, the annotators had the additional task of bracketing them in context and then adding the corresponding semantic classification relations to the bracketed noun - noun pairs. For example, "*((consumer confidence)*EXPERIENCER *survey)*TOPIC" is left-branching. We obtained a bracketing agreement of 87% which was computed as the number of pairs bracketed in the same way by both annotators, over the number of instances classified in the bracketing category considered, by at least one of the judges. For three noun compounds, 33.7% of the instances contained at least one WordNet noun - noun concept that led to an automatic bracketing. From these, 58% were right bracketed, while 68% of the non-WordNet compounds were left bracketed. For example, "*((stock market) boom)*" is automatically left bracketed since "*stock* *market*" is a WordNet concept.

For the two test corpora, we used Lauer's bracketing and prepositional paraphrases annotations. The annotators added the other annotations considered for the training corpora: they disambiguated the noun constituents in isolation (as Lauer provided no context) and tagged the noun - noun pairs with the 35 semantic relations.

The annotators' agreement was measured using the Kappa statistics, one of the most frequently used measure of inter-annotator agreement for classification tasks: $K = \frac{Pr(A)-Pr(E)}{1-Pr(E)}$, where $Pr(A)$ is the proportion of times the annotators agree and $Pr(E)$ is the probability of agreement by chance. The K coefficient is 1 if there is a total agreement among the annotators, and 0 if there is no agreement other than that expected to occur by chance.

Table 2 shows the inter-annotator agreement on the unshuffled training corpora for each semantic interpretation category. We computed the K coefficient only for those instances tagged with one of the 35 semantic relations, respectively 8 prepositional paraphrases. We also computed the number of pairs that were tagged with OTHERS by both annotators for each semantic interpretation relation, over the number of examples classified in that category by at least one of the judges.

In the training corpus, 6.9% of the instances paraphrased with prepositional paraphrases were included in OTHERS category. From these, 4.2% could be paraphrased with other prepositions than those considered by Lauer (eg, "*bus service*" - "*service by bus*"), and 2.7% could not be paraphrased with prepositions (eg, "*daisy flower*").

The K coefficient shows a fair to good level of agreement for the training data on the set of 35 relations, taking into consideration the task difficulty. As Table 2 shows, the level of agreement for the prepositional paraphrases was much higher. All these can be explained by the instructions the annotators received prior to the annotation and by their expertise in lexical semantics.

	Kappa Agreement (1 - 35)		
	2-noun compounds	**3-noun compounds**	**OTHERS**
8 PPs	0.80	NA	91%
35 SRs	0.58	0.69	76%

Table 2
The inter-annotator agreement on the semantic annotation of the noun compounds in the unshuffled training corpora. For the noun compound instances that encoded more than one semantic classification category, the agreement was done on one of the relations only. The agreement on the semantic relations for the three noun compounds was computed on gold bracketing. "NA" means not available.

On the test corpora, the annotation with the set of 35 semantic relations was also done by the two annotators. The disagreement instances were solved by a third judge.

2.3 Distribution of Semantic Relations over the Training and Test Corpora

Although noun compounds are very productive allowing for a fairly large number of possible interpretations, Table 3 shows that a relatively small subset of the 35 semantic relations cover most of the semantic distribution of these constructions on a large open-domain corpus. For example, in the unshuffled 2-noun compounds training corpora there were 25 relations found from the total of 35 relations considered. The most frequently occurring relations were PART-WHOLE, ATTRIBUTE-HOLDER, PURPOSE, LOCATION, TOPIC, and THEME. The semantic relations that did not occur in two noun compounds were KINSHIP, ENTAIL, ACCOMPANIMENT, FREQUENCY, ANTONYMY, PROBABILITY, POSSIBILITY, CERTAINTY, STIMULUS, EXTENT, and PREDICATE.

For three noun compounds, the most frequently occurring semantic relations were ATTRIBUTE-HOLDER, AGENT, TOPIC, and THEME (for left branching), and respectively ATTRIBUTE-HOLDER, AGENT, TEMPORAL, PART-WHOLE, LOCATION, PURPOSE, and THEME (for right branching).

Table 4 shows the mapping between the two sets of semantic classification categories for the unshuffled training corpora.

3 Models for the Interpretation of Two Noun Compounds

The task of noun compound interpretation consists of determining the semantic relations between the noun constituents. In this section we present two main types of learning models: unsupervised and supervised. For both types, the interpretation task is defined as a semantic classification problem. We use two different lists of semantic target categories: the list of 35 semantic relations and the list of 8 prepositional paraphrases and compare our results with those obtained on the same test set by Lauer (Lauer 1995) and Lapata & Keller (Lapata and Keller 2004). Note that Lauer and Lapata & Keller tested their model only on the list of 8 prepositional paraphrases.

3.1 Unsupervised Probabilistic Models

Lauer (Lauer 1995) was the first to devise and test an unsupervised probabilistic model for noun compounds interpretation on Grolier encyclopedia, an 8 million word corpus, based on a set of 8 prepositional paraphrases. His probabilistic model computes the probability of a preposition p given a noun - noun pair $n_1 - n_2$ and finds the most likely prepositional paraphrase

No.	Semantic Relations	N N		N N N			
		%	Example	Left bracket. [%]		Right bracket. [%]	
				$n_1 - n_2$	$n_2 - n_3$	$n_1 - n_3$	$n_2 - n_3$
1	POSSESSION	3.41	"*family estate*"	1.8	2.5	5.12	3.8
2	KINSHIP	0	-	0	0	0	0
3	ATTRIBUTE-HOLDER	8.48	"*quality sound*"	14.1	8.6	16.7	24.4
4	AGENT	4.88	"*crew investigation*"	7.4	21.5	12.8	15.4
5	TEMPORAL	0.93	"*night flight*"	0.6	0.6	14.1	5.1
6	DEPICTION-DEPICTED	0.04	"*image team*"	0	0	0	0
7	PART-WHOLE	16.98	"*girl mouth*"	7.4	9.8	11.5	2.6
8	IS-A (HYPERNYMY)	2.13	"*Dallas city*"	1.2	0	0	0
9	ENTAIL	0	-	0	0	0	0
10	CAUSE	0.04	"*malaria mosquitoes*"	0	0	0	0
11	MAKE/PRODUCE	3.68	"*shoe factory*"	4.3	8.6	0	1.3
12	INSTRUMENT	2.04	"*pump drainage*"	1.8	1.8	0	3.8
13	LOCATION/SPACE	8.14	"*Texas university*"	8.0	6.1	6.4	10.3
14	PURPOSE	11.96	"*migraine drug*"	3.1	4.3	0	14.1
15	SOURCE	1.75	"*olive oil*"	0	1.2	0	0
16	TOPIC	13.07	"*art museum*"	6.1	13.5	3.8	6.4
17	MANNER	0.17	"*style performance*"	2.5	0.6	1.3	0
18	MEANS	0.57	"*bus service*"	0	0	0	0
19	ACCOMPANIMENT	0	"*friends meeting*"	0	0.6	1.3	0
20	EXPERIENCER	0.37	"*disease victim*"	1.2	1.2	1.3	1.3
21	RECIPIENT	0.28	"*worker fatalities*"	6.7	3.1	3.8	0
22	FREQUENCY	0	-	0	0	0	0
23	INFLUENCE	0	-	0	0	0	0
24	ASSOCIATED WITH	0	-	0	0	0	0
25	MEASURE	1.44	"*session day*"	0.6	0.6	0	0
26	SYNONYMY	0	-	0	0	0	0
27	ANTONYMY	0	-	0	0	0	0
28	PROBABILITY	0	-	0	0	0	0
29	POSSIBILITY	0	-	0	0	0	0
30	CERTAINTY	0	-	0	0	0	0
31	THEME	10.99	"*car salesman*"	31.9	10.5	12.1	11.6
32	RESULT	1.64	"*combustion gas*"	1.2	4.9	2.6	0
33	STIMULUS	0	-	0	0	0	0
34	EXTENT	0	-	0	0	0	0
35	PREDICATE	0	-	0	0	0	0
	OTHERS-SR	6.64	"*airmail stamp*"	**3.7**		**7.1**	
Total no. of examples		4504		328		156	

Table 3
The distribution of the semantic relations on the annotated unshuffled training corpora after agreement. The semantic relations for which there was no example given are not encoded by noun compounds.

Sem. rels/ Prep. paraph.	of [%]	for [%]	in [%]	on [%]	at [%]	from [%]	with [%]	about [%]	OTHERS-PP [%]	total
1	96.10	1.94	0	0	0	0	1.94	0	0	154
2	0	0	0	0	0	0	0	0	0	0
3	49.47	0.26	2.09	0.78	0	0	0.26	0	47.12	382
4	68.18	13.63	5.90	2.27	1.81	1.36	0	0	6.81	220
5	45.23	0	33.33	4.76	4.76	0	0	0	11.90	42
6	100	0	0	0	0	0	0	0	0	2
7	71.37	0.65	23.79	0.26	0.26	0.78	1.56	0	1.30	765
8	51.04	0	0	0	0	0	3.12	0	45.83	96
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	100	2
11	62.04	3.61	4.21	0	0	27.10	1.20	1.80	0	166
12	1.08	21.73	2.17	0	0	0	75	0	0	92
13	5.01	0.52	67.81	18.73	7.65	0	0.26	0	0	379
14	9.46	87.94	0.18	2.04	0	0	0.18	0.18	0	539
15	3.79	0	0	0	0	96.20	0	0	0	101
16	18.16	3.56	0.84	9.33	0	0	0.67	67.40	0	589
17	12.5	37.5	12.5	0	0	0	37.5	0	0	8
18	0	0	0	0	0	0	0	0	100	4
19	0	0	0	0	0	0	0	0	0	0
20	94.11	0	5.88	0	0	0	0	0	0	17
21	15.38	76.92	0	0	0	0	7.69	0	0	13
22	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0
25	41.53	4.61	0	0	0	0	0	0	53.84	65
26	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0
31	84.24	6.66	5.85	2.62	0	0	0.40	0.20	0	495
32	10.81	0	1.35	17.56	0	70.27	0	0	0	74
33	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0
total	1869	616	524	175	38	182	103	402	296	4205

Table 4
Mapping between the two sets of semantic classification categories: 8 prepositional paraphrases and 35 semantic relations. The mapping was obtained on the unshuffled training corpora.

$p^* = argmax_p P(p|n_1, n_2)$. However, as Lauer noticed, this model requires a very large training corpus to estimate these probabilities. More recently, Lapata & Keller (Lapata and Keller 2004) replicated the model using the web as training corpus and showed that the best performance was obtained with the trigram model $f(n_1, p, n_2)$. In their approach, they used as count for a given trigram the number of pages returned by Altavista on the trigram corresponding queries. For example, for the test instance "*war stories*", the query was "*stories about war*".

3.2 Supervised Models

The supervised learning models proposed here are centered around two fundamental notions in automatic text understanding: *word sense disambiguation* (WSD) and *lexical specialization* on the general-purpose semantic noun hierarchies offered by WordNet. Each noun in the noun compound is mapped into its corresponding WordNet 2.0 sense determined in context and then classified in its specific WordNet semantic category. So far, we have identified and experimented with the following two features:
`1. Semantic class of head noun` specifies the WordNet sense (synset) of the head noun and implicitly points to all its hypernyms. The NC semantics is heavily influenced by the meaning of the noun constituents. For example: "*GM car*" is a MAKE/PRODUCE relation while "*family car*" is a POSSESSION relation. In case the noun has multiple inheritance, the first semantic class is chosen.
`2. Semantic class of modifier noun` specifies the top semantic class of the WordNet synset. For example "*morning meeting*" - TEMPORAL, while "*business meeting*" - a TOPIC relation.

We present here three supervised models: *Semantic Scattering* (SS) (Moldovan et al 2004), *Iterative Semantic Specialization* (ISS) (Girju et al 2003a), and *Support Vector Machines*. The first two are briefly described below, the third being well known from the machine learning literature (Vapnik 1982).

Semantic Scattering (SS)
The SS model was designed and used by us to semantically classify the genitive constructions and is applicable to noun compounds (Moldovan et al 2004). Essentially, it consists of using a training data set to establish a boundary G^* on the WordNet noun hierarchies such that each feature pair of noun - noun senses f_{ij} on this boundary maps uniquely into one of the 35 semantic relations, and any feature pair above the boundary maps into more that one semantic relations. Due to the specialization property on noun hierarchy, feature pairs below the boundary also map into only one semantic relation. For any new pair of noun - noun senses, the model finds the closest boundary pair, in semantic sense, using a procedure called semantic scattering.

Iterative Semantic Specialization (ISS)
ISS is a multi-classification extension of a binary classification technique initially devised for the PART-WHOLE semantic relation. The iterative semantic specialization method consists of a set of iterative procedures of specialization of the training examples on the WordNet IS-A hierarchy. Thus, after a set of necessary specialization iterations, the method produces specialized examples which through supervised machine learning are transformed into sets of semantic rules for the noun compound interpretation task.

Initially, the training corpus consists of examples that follow the format <*noun1#sense; noun2#sense; target*>, where *target* belongs to the set of classification categories considered. From this initial set of examples an intermediate corpus is created by expanding each example with the corresponding WordNet top semantic classes corresponding to each noun constituent. At this point, the generalized training corpus contains two types of examples: unambiguous, and respectively ambiguous examples. The second situation occurs when the training corpus classifies the same noun - noun pair into more than one semantic category. For example, both relationships <*woman#1, apartment#1,* POSSESSION> and <*woman#1, hand#1,* PART-WHOLE> are mapped into the more general type <*entity#1, entity#1,* POSSESSION/PART-WHOLE>. We recursively specialize these examples to eliminate the ambiguity. By specialization, the semantic class is replaced with the corresponding hyponym for that particular sense, i.e. the concept immediately below in the hierarchy. These steps are repeated until there are no more ambiguous examples.

For the unambiguous examples in the generalized training corpus (those that are classified with a single semantic relation), constraints are determined using cross validation on C4.5.

Support Vector Machines (SVM)
In order to achieve classification in n semantic classes, $n > 2$, we built a binary classifier for each pair of classes (a total of C_n^2 classifiers), and then used a voting procedure to establish the class of a new example. For the experiments with semantic relations, the simplest voting scheme has been chosen; each binary classifier has one vote which is assigned to the class it chooses when it is run. Then the class with the largest number of votes is considered to be the answer.

The software used in these experiments is the package LIBSVM, (*http://www.csie.ntu.edu.tw/~cjlin/libsvm/*) which implements an SVM algorithm. We tested with the radial-based kernel. We experimented with the features generated by the specialization procedure described in the previous supervised models.

The supervised models were trained and tested on both Lauer's data (unshuffled) and random data (shuffled) using the two different lists of semantic classification categories. The results obtained with each model on each test set (*unshuffled*, respectively *shuffled*) are presented in Table 5 using the standard measure of *accuracy* (number of correctly labeled instances over the number of instances in the test set).

We wanted to measure the impact of each basic notion employed in this research, *word sense disambiguation* and *WordNet* IS-A *lexical hierarchy specialization*, and defined two baseline measures. Baseline 1 doesn't take advantage of WSD (sense #1), but it differentiates between unambiguous and ambiguous training examples by specializing the ambiguous ones. In Baseline 2 the noun constituents are tagged with the default sense#1 (no WSD), and the ambiguous examples are not specialized.

List of classif. categ.	**Supervised models**				**Baseline#1 (no WSD, with specializ.)**				**Baseline#2 (no WSD, no specializ.)**			
	SS [%]	**ISS** [%]	**SVM** [%]	**SVM [%] (+PP)**	**SS** [%]	**ISS** [%]	**SVM** [%]	**SVM (+PP)**	**SS** [%]	**ISS** [%]	**SVM** [%]	**SVM [%] (+PP)**
Unshuffled test data												
8 PPs	33.68	39.72	36.26	-	32.35	35.46	33.33	-	32.35	31.91	33.33	-
35 SRs	44.32	37.23	43.53	66.78	34.37	30.17	36.50	61.39	32.03	31.20	20.54	54.25
Shuffled test data												
8 PPs	55.38	50.71	58.07	-	52.74	45.74	54.09	-	48.14	49.29	43.41	-
35 SRs	58.70	43.26	63.91	83.93	50.08	29.08	54.20	77.88	42.56	36.87	27.04	72.59

Table 5
The performance obtained by the supervised models on Lauer's test data (*unshuffled*), and respectively on the random test data (*shuffled*) for the interpretation task. Two sets of semantic classification categories were considered: 8 PPs (8 prepositional paraphrases), and 35 SRs (35 semantic relations)."SVM (+PP)" employes the same feature set as the SVM model plus the corresponding prepositional paraphrase.

The table shows that the supervised models give better results on the list of 35 semantic relations than on the 8 prepositional paraphrases (with the exception of the SS model) on both test sets. This observation is consistent with the initial idea that prepositional paraphrases are more abstract, and thus more ambiguous. Moreover, the comparison with Baseline#1 results shows that word sense disambiguation (WSD) does not represent a very important factor for the noun interpretation as prepositional paraphrases. However, for the classification with 35 semantic relations, the disablement of WSD (sense #1) generates an average drop in accuracy of 7.36% on the unshuffled test set, and respectively of 9.64% on the shuffled test data.

Compared with the WSD feature, the semantic specialization seems to be more important for the noun compound interpretation with prepositional paraphrases, especially for the SVM model on the shuffled test data. Baseline#2 shows an average drop in accuracy of 13.46% for Lauer's test set, and respectively of 17.6% for the shuffled test set. The models most affected by the disablement of both the WSD and specialization features were SS and SVM on both test data sets.

Comparison with Previous Work

On the unshuffled test set, Lauer obtained an accuracy of 40% and Lapata & Keller 55.71%. For the shuffled test set, we replicated Lapata & Keller's experiments (Lapata and Keller 2004) using Google [1] and obtained an accuracy of 46.09%. We formed inflected queries with the patterns they proposed and searched the web. After experimenting with various trigram instances $f(n_1, p, n_2)$, we had the following observations:

1. The order of the constituent nouns in the prepositional paraphrase is important. For example, "war story" (cf. (Lapata and Keller 2004)) can be paraphrased as "story about war" and "story of the war", where the order of the nouns is reversed. However, there are situations in which the order of the nouns remains the same as the one in the noun compound (eg, "*blood vessels*" as "*blood in vessels*" and "*vessels of blood*". For example, 28.19% noun - noun paraphrasable pairs preserved the order in the corresponding prepositional paraphrases. Thus, we tried all plausible alternative queries to cover all possible orderings.

2. Many of the noun compound instances had two or more correct paraphrases. Like Lapata & Keller, in this experiment we considered only the paraphrase with the largest web count provided by the search engine.

3. We manually checked the first five entries generated by Google for each most frequent prepositional paraphrase and noticed that about 38% of them were wrong due to syntactic (eg, POS) and/or semantic ambiguities.

Since we wanted to measure the percentage of syntactic and semantic ambiguities brought by false positive instances, we further tested the probabilistic web-based model on four distinct test sets selected from the Wall Street Journal text collection, each containing 200 noun - noun pairs with different types of ambiguity: in set#1 the noun constituents had only one part of speech and one WordNet sense; in set#2 the nouns had at least two possible parts of speech and were semantically unambiguous, in set#3 the nouns were ambiguous only semantically, and in set#4 they were ambiguous both syntactically and semantically. Table 6 shows that for unambiguous compounds (set#1), the model obtained an accuracy of 34.69%, while for more semantically ambigu-

[1] As Google limits the number of queries to 1,000 per day per computer, we repeated the experiment using 10 computers for a number of days. Although Keller & Lapata used Altavista for the interpretation of two noun compounds, they showed that there is almost no difference between the correlations achieved using Google and Altavista counts.

ous compounds it obtained an accuracy of about 50% (sets #3 and #4). This shows that for more semantically ambiguous noun - noun pairs, the web-based probabilistic model introduces a significant number of false positives.

Test set	**Ambiguity**		**Accuracy [%]**
	Syntactic (POS)	**Semantic (WSD)**	
Set#1	No	No	34.69%
Set#2	Yes	No	33.01%
Set#3	No	Yes	50.82%
Set#4	Yes	Yes	44.46%

Table 6
Experimental results with Lapata & Keller's web-based unsupervised interpretation model on different types of test sets. "No" means not ambiguous and "Yes" means ambiguous.

4 Models for the Interpretation of Three Noun Compounds

The interpretation of three noun compounds consists of two phases: the bracketing and the automatic annotation with semantic categories. In this section we present experimental results with unsupervised probabilistic and supervised models on bracketing and semantic annotation of 3-noun compounds. The results are drawn from two test sets: Lauer's 244 test data (*unshuffled*), and a randomly selected set of 244 noun compound instances (*shuffled*). For the semantic annotation we use the list of 35 semantic relations proposed in Section 2.

4.1 Unsupervised Probabilistic Models for the Bracketing of Three Noun Compounds

The task of noun compound bracketing is defined as follows: given a three-word noun compound $n_1\ n_2\ n_3$, if $(n_1\ n_2)$ is the most correct bracketing of the noun sequence, then the structure is $((n_1\ n_2)\ n_3)$, otherwise the correct structure is $(n_1\ (n_2\ n_3))$.

Most of the unsupervised probabilistic approaches to noun compound bracketing ((Resnik 1993), (Lauer 1995), (Lapata and Keller 2004)) are based on two models: *adjacency*, and respectively *dependency model* (cf. (Lauer 1995)). The *adjacency model* compares frequencies of $(n_1\ n_2)$ to $(n_2\ n_3)$. The *dependency model* compares the probabilities of occurrence of $(n_1\ n_2)$ to $(n_1\ n_3)$, ignoring previous occurrences of $(n_2\ n_3)$. Lauer estimated the frequencies of each possible bracketing on Grolier encyclopedia based on a taxonomy or thesaurus. In equations (1) and (2), for example, t_1, t_2, and t_3 represent thesaurus conceptual categories and w_i are noun members of these categories. The probability

$P(t_1 \to t_2)$ denotes the modification of a noun in category t_2 by a noun in category t_1.

$$R_{adj} = \frac{\sum_{ti \in cats(w_i)} P(t_1 \to t_2)}{\sum_{ti \in cats(w_i)} P(t_2 \to t_3)} \quad (1)$$

$$R_{dep} = \frac{\sum_{ti \in cats(w_i)} P(t_1 \to t_2) P(t_2 \to t_3)}{\sum_{ti \in cats(w_i)} P(t_1 \to t_3) P(t_2 \to t_3)} \quad (2)$$

Like Lapata& Keller, we also experimented with both adjacency and dependency web-based models on the shuffled test set using lexical items rather than semantic categories.

4.2 Supervised Model for the Bracketing and Semantic Annotation of Three Noun Compounds

For the bracketing subtask, we experimented with a list of fifteen linguistic features employed in the C5.0 decision tree model. For each of the three nouns in a compound, the following five features were computed based on the WordNet sense of each noun constituent determined in context:

1. `WordNet derivationally related form` specifies if that sense of the noun is related to a verb in WordNet. Example: "*coffee maker industry*", where the correct sense in this case *maker#3* is related to the verb *to make#6*.
2. `WordNet top semantic class of the noun`. Example: "*coffee maker industry*", where *maker#3* is a {*group, grouping*}#1.
3. `WordNet second top semantic class of noun`. Example: "*coffee maker industry*", where *maker#3* is a *social_group*#1.
4. `WordNet third top semantic class of noun`. Example: "*coffee maker industry*", where *maker#3* is *organization#1*.
5. `Nominalization` indicates if the noun is a nominalization or not based on the NomLex dictionary of nominalizations (Macleod et al 1998). We also consider nominalizations those nouns that could not be found in NomLex but are *events* or *actions* in WordNet. Example: "*coffee maker industry*", where *maker* is a nominalization.

For the semantic annotation subtask we used the same feature set at which we added the bracketing information as a new feature. We experimented with three different target classification sets: semantic relation SR#1 defined here as the semantic relation between $n_1 n_2$ (if left branching), and respectively between $n_1 n_3$ (if right branching); semantic relation SR#2 defined as the semantic relation between $n_2 n_3$ (if left or right branching)); and combinations of semantic relation #2 with semantic relation #1 as feature, and semantic relation #1 with semantic relation #2 as feature. Consider the following examples:

(4) *((debt#2 reduction#1)/*THEME*/r exercise#3)* /PURPOSE,

(5) *(morning#1 (package#2 sort#4)/*THEME*/r)* /TEMPORAL

In (4) the semantic relations were added on a left-branching structure, and in (5) on a right-branching one. The tag "*r*" indicates that the nouns are in *reverse* order.

4.3 Experimental Results and Observations

For the bracketing task, we compared the results obtained with both the supervised and the unsupervised probabilistic models. For the semantic annotation task we used only the C5.0 decision tree model. All models were tested on each of the two test sets: Lauer's data set (unshuffled) and the shuffled set.

On the unshuffled test set, Lauer obtained an accuracy of 80.70% and Lapata & Keller (Lapata and Keller 2004) an accuracy of 78.68% with the dependency model (77.86% with the adjacency model respectively). For the shuffled test set, we replicated Lapata & Keller's bracketing experiments using again inflected queries on Google, and obtained an accuracy of 77.36% with the dependency model, and 73.45% with the adjacency model respectively. The results obtained with each model on each test set are presented in Table 7 and are compared against a baseline (no WSD).

Comparison with previous work
According to our knowledge, all solutions proposed before to the automatic interpretation of 3-noun compounds focused only the bracketing problem. As mentioned before, most of these approaches are probabilistic and are based on the assumption that the probability of occurrence of a pair of nouns is independent of the third noun, which most of the time is unrealistic and leads to errors.

Unlike previous work, we focus on a set of semantic features employed in a supervised machine learning model. Instead of focusing on noun compounds in isolation, our model brackets them in context through the use of the WSD feature.

5 Compound Nominals Interpretation applied to Question Answering

A powerful method of answering more difficult questions is to associate to each question the semantic relation that reflects the meaning of that question

Learning models		**Semantic Annotation Task (semantic relations cf. bracketing)**		**Bracketing Task**
		SR#1	**SR#2**	
		n_1n_2 **(left),** n_1n_3 **(right)**	n_2n_3 **(left and right)**	
Supervised	**Unshuffled test data**			
	C5.0	45.10%	37.10%	73.10%
	Baseline (no WSD)	36.21%	31.01%	72.80%
	C5.0 (+SR#1)	-	26.40%	-
	C5.0 (+SR#2)	34.30%	-	-
	Shuffled test data			
	C5.0	50.50%	50.60%	83.10%
	Baseline (no WSD)	42.08%	40.20%	74.40%
	C5.0 (+SR#1)	-	53%	-
	C5.0 (+SR#2)	54.08%	-	-
Unsupervised probabilistic (Lapata & Keller's web-based model)	**Unshuffled test data**			
	Adjacency	-	-	77.86%
	Dependency	-	-	78.68%
	Shuffled test data			
	Adjacency	-	-	73.45%
	Dependency	-	-	77.36%

Table 7
The performance obtained by the supervised and unsupervised probabilistic models on the two test data sets for the interpretation task. "C5.0 (+SR#i)" means the supervised model applied to the basic feature set plus SR#i as feature.

and then search for that semantic relation over the candidates of semantically tagged paragraphs (Moldovan et al 2002). Here are some examples.

Q. Where have nuclear incidents occurred? From the question stem word *where*, we know the question asks for a LOCATION which is found in the complex nominal "*Three Mile Island*" -LOCATION of the sentence "The Three Mile Island nuclear incident caused a DOE policy crisis", leading to the correct answer "*Three Mile Island*".

Q. What is the source of radiation in nuclear waste? The answer is found in the text "The uranium isotope radiation in nuclear waste is toxic in small doses." The NC "((*uranium isotope) radiation*)" contains a SOURCE semantic relation meaning that "uranium isotope" is the source of "radiation" which matches the question semantic, thus the answer is "uranium isotope".

Q. What did the factory in Howell Michigan make? The verb *make* tells us to look for a MAKE/PRODUCE relation which is found in the complex nominal "*car factory*"-MAKE/PRODUCE of the text: "The car factory in Howell Michigan closed on Dec 22, 1991" which leads to answer *car*.

Q. *What was the purpose of the stolen equipment?* The answer is found in the complex nominal "*dive equipment*"-PURPOSE of the sentence: "The dive equipment was stolen between 10 pm and midnight on Tuesday", leading to the correct answer "*dive*".

6 Discussion

Our approach to noun compound interpretation is novel in several ways. The semantic interpretation problem is tackled for both two and three noun compounds. We provide empirical observations on the distribution of the meaning of noun compounds on fairly large corpora by performing various experiments with human judgments on two state-of-the-art semantic classification lists. A mapping between the two classification sets based on the noun compounds distribution on the corpora is also provided. For both the bracketing and semantic annotation tasks, the paper presents experimental results with supervised learning models based on linguistic features and compares them against state-of-the-art probabilistic approaches and against the optimal performance obtained by two human annotators.

Our supervised models use an iterative semantic specialization method that allows us to go deeper into the semantic complexity of noun compounds. According to our knowledge, the system is the only domain independent noun compound interpretation tool that uses word sense disambiguation and WordNet IS-A specializations. One other symbolic system, SENS (Vanderwende 1995) makes some use of IS-A generalizations, but considers only the first sense of the noun constituents in WordNet. The current state-of-the-art systems in automatic detection of semantic roles (Gildea and Jurafsky 2002) that are probabilistic-based have also tried to use lexico-semantic hierarchies, such as WordNet, to generalize from lexical noun features. However, they also rely on the first sense listed for each noun occurring in the training data. Other approaches, such as Resnik's conceptual association algorithm (Resnik 1996), attempt to automatically detect semantic roles based on semantic similarity measures applied to large lexical hierarchies, such as WordNet. However, not many of these attempts make use of lexico-semantic hierarchies for generalization, due to the unavailability of word sense disambiguation tools. For example, Resnik and Hearst (Resnik and Hearst 1993), and Brill and Resnik (Brill and Resnik 1994) have used the conceptual association algorithm to solve the prepositional phrase attachment problem. However, the similarities computed on WordNet for various noun-noun or verb-noun pairs linked by a preposition were not sufficient for the attachment task, as the pairs were not linked hierarchically.

The system presented here makes use of an existing lexical resource, WordNet, that contains general purpose information that can be successfully used in domain independent applications. Moreover, the system maps the disambiguated noun constituents into the WordNet noun hierarchies. The system is also unique in the sense that it uses a specialization procedure on the WordNet noun hierarchies in order to generate the best semantic constraints for the noun compound interpretation.

One main drawback of the approach is the use of supervised models that require a large annotated corpus. Another drawback is the heavy reliance on WordNet which has been criticized by some. As we have demonstrated, the impact of the WSD is considerable, however, the current state-of-the-art in WSD is not satisfactory yet.

7 Acknowledgments

We like to thank Mark Lauer for providing the two noun and three noun test sets and to Frank Keller and Mirella Lapata for valuable comments.

References

[Brill and Resnik 1994] Eric Brill and Philip Resnik. A rule based approach to PP attachment disambiguation. In *Proceedings of 15th International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japan.

[Charniak 2000] Eugene Charniak. 2000. A maximum-entropy-inspired parser. In Proceedings of the *North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, Seattle, WA.

[Downing 1977] P. Downing. On the creation and use of English compound nouns. *Language*, 53(4), 810-842.

[Gildea and Jurafsky 2002] D. Gildea and D. Jurafsky. Automatic Labeling of Semantic Roles. In *Computational Linguistics* 28:3, 245-288.

[Finin 1980] T. Finin. The semantic interpretation of compound nominals. University of Illinois at Urbana-Champaign.

[Girju et al 2003a] R. Girju, A. Badulescu, and D. Moldovan. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In *Computational Linguistics*, accepted for publication with reviews.

[Moldovan and Girju 2003] R. Girju and D. Moldovan. Knowledge Discovery from Text. *Tutorial Proceedings*at the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), Sapporo, Japan.

[Jespersen 1954] O. Jespersen. A modern English grammar on historical principles. In George Allen & Unwin Ltd. (Eds.), London, 1909-49.

[Lapata 2000] M. Lapata M. The Automatic Interpretation of Nominalizations. In the *Proceedings of AAAI*, pages 716-721.

[Lapata and Keller 2004] M. Lapata and F. Keller. The Web as a Baseline:

Evaluating the Performance of Unsupervised Web-based Models for a Range of NLP Tasks In the proceedings of the *Human Language Technology* conference (HLT/NAACL 2004), Boston, MA.

[Lauer 1995] M. Lauer. Designing Statistical Language Learners: Experiments on Noun Compounds. In *PhD thesis*, Macquarie University, Australia.

[Levi 1978] J. Levi. The syntax and semantics of complex nominals. In NY: Academic Press.

[Macleod et al 1998] C. Macleod, R. Grishman, A. Meyers, L. Barrett, and R. Reeves. 1998. Nomlex: A lexicon of nominalizations. *8th International Congress of the European Association for Lexicography*, Belgium.

[Moldovan et al 2002] D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badulescu, and O. Bolohan. LCC Tools for Question Answering. In Proceedings of TREC-2002.

[Moldovan et al 2004] D. Moldovan, A. Badulescu, M. Tatu, D. Antohe, and R. Girju. Models for the Semantic Classification of Non-nominalized Noun Phrases. In the proceedings of the *HLT/NAACL 2004* - Computational Lexical Semantics workshop, eds. D. Moldovan and R. Girju, Boston.

[Pustejovsky et al. 1993] J. Pustejovsky, S. Bergler, and P. Anick. Lexical semantic techniques for corpus analysis. In *Computational Linguistics*, 19(2).

[Resnik and Hearst 1993] P. Resnik and M. Hearst. Structural Ambiguity and Conceptual Relations. In the *Proceedings of 1st Workshop on Very Large Corpora*.

[Resnik 1993] P. Resnik. Selectional Constraints: An Information-Theoretic Model and its Computational In *Doctoral Dissertation*, Department of Computer and Information Science, University of Pennsylvania.

[Resnik 1996] P. Resnik. Realization. In *Cognition*, 61:127-159.

[Quinlan 1993] J. Ross Quinlan. C4.5: Programs for Machine Learning. *Morgan Kaufmann.*

[Rosario and Hearst 2001] B. Rosario and M. Hearst. Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy. In the *Proceedings of EMNLP '01*, Pittsburgh, PA.

[Rosario et al 2002] B. Rosario, M. Hearst, and C. Fillmore. The Descent of Hierarchy, and Selection in Relational Semantics. In the *Proceedings of ACL '02*, Philadelphia, PA.

[Siegel and Castellan 1988] S. Siegel and N.J. Castellan. 1988. *Non Parametric Statistics for the behavioral science.* New York: McGraw-Hill.

[Vanderwende 1995] L. Vanderwende. The Analysis of Noun Sequences using Semantic Information Extracted from On-Line Dictionaries. *Ph.D. thesis*, Georgetown University.

[Vapnik 1982] V. Vapnik. Estimation of Dependences Based on Empirical Data. In *Springer Verlag.*

[XWN 2003] eXtended WordNet. http://xwn.hltri.utdallas.edu